Method and apparatus for automatic detection of data types for data type dependent processing

The invention relates to a method and an apparatus for the
5 classification, organization and structuring of different
types of data, which can be used e.g. for data sorting, data
storage or data retrieval.

10 Background

The capacity of digital storage media like hard disks or
rewritable optical disks for personal recording of video and
other data grows continuously. This results in new concepts
15 like e.g. the so-called home server, which is a central
storage device with large capacity for recording any kind of
data within the home. Such applications also require new
ways to organize the recorded data, search for content and
access specific recordings.
20

For this purpose data about data, often referred to as
metadata, can be used. Various industry groups and standard
bodies have been developing metadata standards for different
purposes and applications. In multimedia applications,
25 metadata typically are data about audiovisual (AV) data,
these AV data often being called 'essence'. However, a Data
Base Management System (DBMS) that shall be able to handle
data of various data types correctly requires a definition
of data types, and a method to distinguish between them.
30

Invention

The invention is based on the recognition of the facts
35 described in the following:
In devices providing a DBMS for handling of incoming data,

including incoming metadata, it is necessary to classify said incoming data, and especially incoming metadata, since different processing is necessary for different kinds of metadata. For example, a text query is not suitable for

5   metadata containing a picture in the well-known Graphics Interchange Format (GIF).

The problem to be solved by the invention is to classify the data automatically, such that a DBMS can utilize the result

10   of the classification for correct data handling. This problem is solved by the method disclosed in claim 1 and by the apparatus disclosed in claim 5. The output of such apparatus may be directed towards e.g. a DBMS.

15   According to the invention, Metadata can be defined as data sets consisting of two parts, namely a first part being a link, the link pointing to a reference data set, and a second part being any data referring to said link. In the following, said first part is referred to as MD_LINK, and

20   said second part is referred to as MD_LOAD. Any data item that does not contain at least one MD_LINK and a related MD_LOAD is defined to be Essence. Metadata often occur together with other Metadata or Essence, combined in a logical entity like e.g. a file on a hard disc. Such mixture

25   of different kinds of Essence and Metadata is in the following called 'Container'. Popular examples for such Containers are Hypertext Markup Language (HTML) files, or Portable Document Format (PDF) files.

30   Further, according to the invention there is another type of classification possible. Data may require interpretation through the device before they can be used. In this case the data are defined to be Physical Data, if the device has a method for interpretation defined, otherwise Abstract Data.

35   If e.g. a picture is stored in GIF format, and the device can interpret GIF format and display it as a picture, it is

classified as Physical Data. If the device cannot interpret
GIF format, the data is classified as Abstract Data. Further
examples for Abstract Data are text files, and other files
that cannot be interpreted through the device.

The previously defined two types of classification are not
exclusive, but complementing each other. Further, the
described classification of data is not absolute, but system
dependent, and therefore only locally relevant.

Advantageously, this classification allows the device to
handle different data types correctly, differ between
Metadata, Essence, Container, Physical Data and Abstract
Data, and thus permit a generalized access method upon said
data types. With this knowledge, the device can decide e.g.
which type of data-query to use, how to interpret data, and
if some data can be disregarded for a certain query.

Advantageous additional embodiments of the invention are
disclosed in the following text, and in the respective
dependent claims.

Drawings

Exemplary embodiments of the invention are described with
reference to the accompanying drawings, which show in:

Fig. 1   the two systems, or dimensions, of data
            classification;

Fig. 2   an example for a Container containing Essence and
            Metadata;

Fig. 3   an example for Abstract Metadata;

Fig. 4   an example for Physical Metadata; and

Fig. 5   an exemplary flow-chart for the method according
         to the invention.

5

Exemplary embodiments

According to the invention, the two types, or systems, of
10   classification can be understood as two dimensions, as shown
in Figure 1. A data item may either be Essence E or Metadata
M, and either Physical Data PD or Abstract Data AD.
Therefore the possible data types are Physical Essence PE,
Physical Metadata PM, Abstract Essence AE or Abstract
15   Metadata AM. Further, a data item may also be a Container C,
if it contains other data items.

The classification of data is not absolute, but subjective
from the sight of the device, and therefore only relevant
20   within a system, e.g. DBMS. It may happen that e.g. one
system can interpret a link while another system cannot
interpret the same link. Therefore it may happen that e.g.
one system classifies certain data as Metadata, consisting
of MD_LOAD and MD_LINK, while another system classifies the
25   same data as Essence because it cannot interpret the link.
Another example is that e.g. one system can reproduce an
MPEG audio layer 3, or MP3, coded file, while another system
cannot interpret the MP3 format. In this case the first
system classifies an MP3 coded file as Physical Data, but
30   the second system classifies the same file as Abstract Data.

Text is to be regarded as Abstract Data, because text is
always a format for saving data. Formatted text can
represent a direct physical representation of data, e.g. the
35   PDF format. The format information represents only support
information, i.e. if format information is extracted from a

PDF file, the pure text being the main information will remain. If the text is extracted, the main information will be lost. Due to the fact that the text represents the main information, also formatted text will be regarded as

5   Abstract Data.

A device as disclosed in claim 5 will execute the following procedure when receiving data on its input:

        If the data contain more than one data item, the output
10.   may be: "Data is a Container". More details are given
        below. Classification may stop here, or may be extended
        to some, or all, leaves of the hierarchically
        structured data tree within the Container.
        If data are Metadata, the output may be: "Data are
15   Metadata".
        Otherwise the output may be "Data are Essence".
        If data are Physical Data, an additional output may be
        "Data are Physical Data".
        Otherwise, if data are Abstract Data, an additional
20   output may be "Data are Abstract Data".
        Advantageously the device can detect and output the
        type of Physical Data, e.g. "Data is a color picture
        (24bit) with the resolution x=200 pixels and
        y=400 pixels".
25   If the data format is unknown to the device, and
        therefore the device is not able to classify the data
        as Container, Metadata, Essence, Abstract Data or
        Physical Data, the output may be any default-type
        output, e.g. "Data type is unknown" or "Data are
30   Essence and Abstract Data".

Additionally, it is helpful if the device detects whether data is text or not:

        If data is Abstract Data and text, the output may be
35   additionally "Data is Text".
This may be implemented by searching for known words, e.g.

from an electronic dictionary, or searching for groups of
characters separated by blanks.

5   If the input data is a Container, an additional output may
     be "Data is a Container, i.e. more metadata or essence are
     contained". Optionally, precise details can be included:
     "The Container CONTAINS at least 1 Metadata and 1 Essence",
     or "The Container CONTAINS no Metadata at all" or even "The
     Container CONTAINS exactly N Metadata items", with N being
10   the amount of Metadata contained in the Container.

     If the device can detect the format of the analyzed data, it
     may output it additionally: "Data format is X". 'X' is the
     format. Examples for 'X' can be e.g. 'HTML' or 'JPEG'.
15

     Figure 2 shows an example for a data file containing a
     combination of Essence and Metadata in the well-known HTML
     format. In the following, the classification of all elements
     according to the invention is described.
20

     First the device detects that the first line is <html>, and
     that therefore the data file should be HTML formatted. It is
     assumed that the device can interpret the HTML format, and
     therefore interprets items with "href" attributes in HTML
25   files as links. Since HTML formatted files usually contain a
     hierarchical structure, the leaf elements of the hierarchy
     tree are analyzed first. The first element from Fig.2
       <title>This is the title</title>
     is classified as Essence because there is no link attached
30   to the element.

     The element
       W3C HOME
     is classified as Metadata, with the string "W3C HOME" being
35   the Essence, or MD_LOAD, and the string
     "href=http://w3c.org" being the related link, or MD_LINK.

The next leaf element
 <p>This is a paragraph</p>
contains no link and is therefore classified as Essence.

The next leaf element
 <img src="image.gif">
is also classified as Essence because it is only a link,
i.e. it contains no MD_LINK with related MD_LOAD. Therefore
it cannot be Metadata. The purpose of this link is to
reference further Essence, namely the picture data.

When all elements of the first level of hierarchy are
analyzed, the next level is investigated. The element
 <head>
  <title>This is the title</title>
 </head>
is classified as Essence because it contains no link, but
only one element, the element being Essence.

The element

  <img src="image.gif">

is classified as Metadata, with <img src="image.gif"> being
the MD_LOAD part and the "href" attribute being the related
link.

The next element
 <body>
   ...
 </body>
is classified as Container because it groups together
Metadata items and Essence items.

Finally, the element

```
<html>
   ...
</html>
```

is also classified as Container. It groups together an
Essence element, namely the <head> element, and a Container,
namely the <body> element.

Figure 3 shows an example for Abstract Metadata. Several
data items 3R,3M are grouped in a data unit 3C. The data
unit 3C could be e.g. an HTML file. For one of said data
items the device has detected that it contains a link 3L,
symbolized by the cursor switching from an arrow to a hand
when pointing to the text 3E. Since the text 3E and the link
3L belong together, and the text 3E is Essence, they form a
Metadata item 3M, and the link 3L is a Metadata link
pointing to a reference 3REF outside the data unit 3C. Since
the Essence 3E of the Metadata item 3M is text, and text is
Abstract Data, the Metadata item 3M is an Abstract Metadata
item. Remaining data items 3R within the data unit 3C are
any text and a picture. The data unit 3C is a Container,
since it contains at least one Metadata item 3M and other,
remaining data items 3R.

Figure 4 shows an example for Physical Metadata. Several
data items 4R,4M are contained in a data unit 4C, the unit
4C being e.g. an HTML file. In this case, the device has
detected that the picture 4E is associated to a link 4L,
symbolized by the cursor switching from an arrow to a hand.
The link 4L is pointing to a reference 4REF outside the data
unit 4C. Since the picture 4E and the link 4L belong
together, they form a Metadata item 4M, with the picture 4E
being the Essence of this Metadata. Said Essence 4E is e.g.
a JPEG formatted picture, and in the HTML file it may be
referenced e.g. as <img src=Anton.jpg width=108 height=73>.
Since the device can display it, it is Physical Data, and
the Metadata item 4M is Physical Metadata. The data unit 4C

is a container, because it contains at least one Metadata
item 4M and other items 4R.

Figure 5 shows an exemplary flow chart of the inventive
5   method. The purpose of the invention is to classify
different types of incoming data IN. The incoming data IN
are being analyzed, and a first decision block D1 decides
whether the format of the incoming data can be detected. If
not, 'Unknown' is indicated as an output, and the
10   classification finishes at an end state EX. If the format is
known, e.g. HTML, then a second decision block D2 may decide
if the incoming data contains unclassified elements. If the
answer is 'Yes', the next unclassified data item is picked
and forwarded to a third decision block D3. This decision
15   block D3 may decide whether said data item is a Container C,
Metadata M or Essence E. The decision is 'Container' if the
data item contains another data item already classified as
Metadata. The decision is 'Metadata' if the data item
contains a link with essence relating to that link. In all
20   other cases the decision is 'Essence'. The decision made in
the third decision block D3 is indicated at the output. If
the analyzed data item is a Container C, then the procedure
returns to the second decision block D2 again, otherwise a
fourth decision block D4 is entered. Said fourth decision
25   block D4 decides whether the device can interpret the data
item, such that it may disclose further information to the
user, e.g. a displayable picture. If the answer is 'Yes', it
is indicated at the output that said data item is Physical
Data PD, otherwise Abstract Data AD. In the case of said
30   data item being Physical Data PD, format detection may have
been done implicitly in said fourth decision block D4. Then
a fifth decision block D5 may detect format details and
decide whether the detected format shall be indicated, and
if so, the format F1,…,F3 may be indicated at the output. In
35   the case of said data item being Abstract Data AD, a sixth
decision block D6 may decide if the data contains text. If

so, this is indicated at the output. If the data item is
Abstract Data AD and not text, no further indication is
generated. Then the procedure is repeated from the second
decision block D2 that decides if further unclassified
5 elements are contained. If this is not the case, then the
data item has been classified completely and the end state
EX is entered. This embodiment of the invention analyzes all
hierarchy levels and leaf elements of Containers, but other
embodiments may analyze only some hierarchy levels or leaf
10 elements of Containers.

Advantageously, the described method for data classification
can be used in devices for data sorting, data storage e.g.
DBMS, or data retrieval e.g. browsers. The described method
15 can be used when different classes of data require different
processing, e.g. different search algorithms, different
storage methods or areas, different compression methods or
different presentation methods.

20 The invention can be implemented in a separate device, which
will classify incoming data with respect to its format,
content, and relation to other data, e.g. link, and which
provides information about data. This information is
especially necessary when it is to recognize, whether these
25 data contain links or these data need special query-methods.

The device can be part of another device or can be realized
as hardware or software, e.g. as an application or a plug-in
in a PC. Further, it can be updated, e.g. via the Internet
30 or via other sources, so that more and more formats can be
recognized, thus this device will update itself and get more
and more efficient.